Emotion Classification Based on Face Images

Zhenyu Wei, Siqi Bian (Susie), Wenyu Li
University of California, Davis

June 6, 2016

Abstract

The objective of the project is to predict the corresponding emotion when given a new face image based on unsupervised and supervised machine learning methods. The data used contains 610 images which are encoded six levels of emotion labels: 0=neutral, 1=anger, 2=disgust, 3=happy, 4=sadness and 5=surprise. Data pre-processing procedure is applied to be convenient for the analysis, which includes face images gray-scale converting, detecting and dimensional reduction. For unsupervised part, clustering analysis is applied for one and two persons. It's found that clustering to 4 groups can achieve the best result. As for supervised machine learning, linear and nonlinear SVM, logistic regression, KNN and other three algorithms are applied to fit eight different models, tuning parameters with 5-fold stratified cross validation on the training data. In the end, a pooling method is conducted to improve the prediction ability combining the best two models which have the smallest CV error rate. It's concluded that our methods can obtain relatively high prediction ability.

Key words: Face, emotions, classification, supervised, unsupervised, machine learning, prediction ability.

1 Introduction

People usually have different facial expressions according to their emotions. Human can easily recognize the emotions on people's face. Here we want to achieve the goal of emotion recognition on face images based on machine. We used a subset of the Cohn-Kanade dataset[1], with 610 images with 640*400 pixels in total. Each image has a label with six levels of emotions: 0=neutral, 1=anger, 2=disgust, 3=happy, 4=sadness and 5=surprise. We would like to explore the Unsupervised Machine Learning methods to distinguish the different emotions given a set of different emotional images of one or two people. Furthermore, apply some Supervised Machine Learning methods to recognize precisely its corresponding emotion given a new image.

In order to classify the images well, we convert all of images to grayscale, detect the face, and crop them to only keep the faces as a pre-processing procedure.

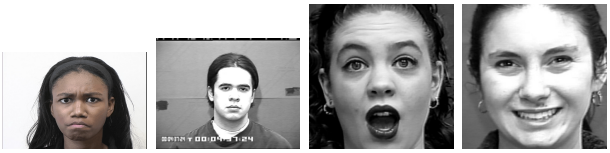


Figure 1: Original and Pre-processed Images

2 Clustering (Unsupervised)

2.1 Clustering for one person's emotion

We have used different linkage methods in agglomerative hierarchical procedures, like average linkage, complete linkage, single linkage methods, etc. We could easily find that, in average, linkage method, the results of clustering are reasonable, because each clustering can represent one emotion. There are 4 kinds of emotions: surprise, sadness, neutral, and happy. We found that different linkage methods show the same clustering results.

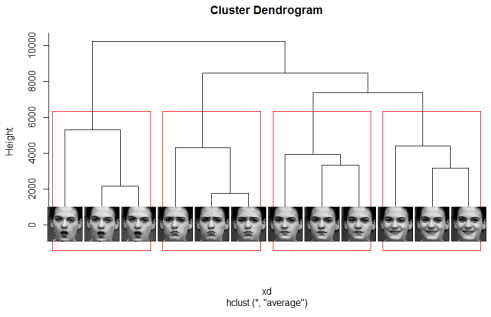


Figure 2: Clustering for one person's emotion

Moreover, K-means method and K-medoids methods also show the same clustering result, and the number K of clusters is determined by mean of silhouette value[2].

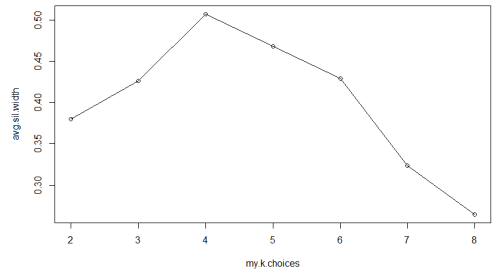


Figure 3: silhouette values v.s. cluster numbers K

The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. From the plot above, we could easily find that K = 4 is the best choice.

When we perform clustering analysis on another person's face images, the clustering results is also good. There are also 4 kinds of emotions: happy, fear, neutral, contempt.

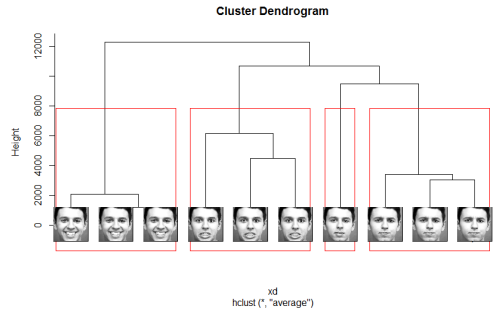


Figure 4: Clustering for one person's emotion

2.2 Clustering for two people's emotion

Then we try to perform clustering analysis based on two person's face images. We find that clustering method could also perform a good result.

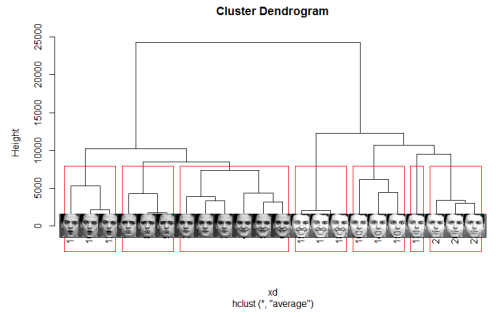


Figure 5: Clustering for two persons' emotion

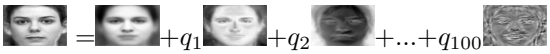
Each cluster still represents one kind of emotion of one person. However, when we add lots of different people's face image, the feature of each clusters would become no obvious.

3 Supervised Machine Learning

3.1 Dimensional Reduction

Since the 3-dimensions 610×350×350 data cannot be used on fitting the classification models, the dimensional reduction is necessary here. The steps are as following:
Considering we have 610 face images with 350×350 pixels. We can view each image as 350×350, which is 122500 dimensions data and denote it as p. We have 610 face images, denote it as n and so our original face data matrix can be denoted $X_{n\times p}$. Transforming the data to $Y = \frac{1}{\sqrt{n-1}}(X - 1\mu')$, where $\mu' = \frac{1}{n}1'X$. Since the sample covariance matrix $Y'Y$ has 122500×122500 dimensions, we cannot directly compute eigenvalues and eigenvectors of it. Based on the knowledge of SVD and PCA[3], the eigenvectors of $Y'Y$ is $\phi = Y'u$, where u is the eigenvectors of YY'. (Usually, we have to normalize ϕ). Select m = 100 principle components, which denotes as $\Phi_m = (\phi_1, ..., \phi_m)$. Then for each face image x, the score of its component is

$$\hat{Y}_{m\times 1} = \Phi'_m(x - \mu)$$

Therefore, $\hat{x} = \Phi_m\hat{Y} + \mu$. Since m < (n,p), we can project our data onto a lower dimension space. Then we could use it for recognition of face images.
For example, the 8th image is  We can reconstruct this image with 100 PCs:


3.2 Model Selection

To get the prediction of labels based on different face images, we decide to apply different models and choose the models with the best prediction ability. Since the sample sizes under each emotion are different (from label 0 to 5, the sample sizes are respectively 327, 45, 59, 69, 25, 81), we used stratified random sampling to split our dataset into two parts: train set (80%) and test set (20%).
We fit eight different models and tuned the parameters with a 5-fold stratified cross validation on the training set: linear SVM (tuned C and gamma), radial basis function kernel SVM (tuned C and gamma), 3-degree polynomial SVM (tuned C and gamma), logistic regression (tuned C), K Nearest Neighbors (tuned K), Naive Bayes, Random Forest tree (tuned the number of trees in the forest) and linear discriminant analysis. Then we got the average of the error rates from the 5 fold CV for each of these models.

The following table shows the result of the chosen parameters which can obtain the smallest cross validation error rate.

Model	Error
Linear SVM	0.166
Rbf SVM	0.217
Poly SVM	0.463
Logistic Regression	0.143
KNN	0.438
Naive Bayes	0.296
RFT	0.336
LDA	0.174

Table 1: Error Rate for Different Algorithms

From the table, we can see the linear SVM (C = 0.01) and logistic regression (C = 0.1) algorithm have the best predictive ability among these different methods, since they have the smallest cross validation error rate. To get a better prediction, we do a pooling process by constructing an ensemble learning on these two best methods. The detailed analysis is in the following part.

3.3 Pooling and Prediction

Pooling method is a general approach to combine the information from different sources and probability distributions[2]. Based on the linear SVM (C = 0.01) and logistic regression (C = 0.1), We'd like to ensemble them in order to balance out their individual weaknesses and obtain a better prediction because they are almost equally well performing when training the data. Then the aggregate probability is

$$P(y_i|c_j) = \sum_{k=1}^{k} \alpha_k P(y_i|c_j)$$

where K is the number of models (Here K = 2). $P(y_i|c_j)$ means the predicted probability assigned by the model K to the response y_i occurring in the multiclass c_j. α_k is the weights:

$$\alpha_k = log\frac{1 - error_k}{error_k}$$

Then normalized α_k to sum to one and plug into the following functions.
Since it is a multiclass problem with 6 labels in total, we used One-Vs-One method by constructing one classifier per pair of classes during the procedure of getting the final probability.
After constructing the new pooled model, we obtained the normalized weights $\alpha_{linearsvm} = 0.475$ and $\alpha_{logistic} = 0.525$ by using the errors obtaining from the training set. Then we applied the pooling model, the linear SVM (C = 0.01) and the logistic regression (C = 0.1) on the new data from the 20% test data-set. Compare the results of the pooling model and the results of those two models:

	Pooling	linear SVM	logistic
Accuracy	0.861	0.844	0.836

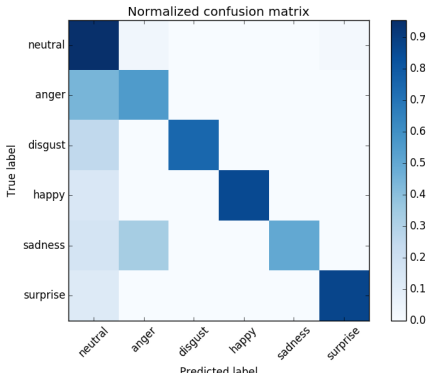
Therefore, the pooling method has a higher predicted accuracy score and obviously improves the performance.

4 Conclusion and Discussion

The following is the table of accuracy score on each labels using pooling model and the best two models.

	Pooling	linear SVM	logistic
neutral	0.95	1.00	0.94
anger	0.56	0.44	0.44
disgust	0.75	0.75	0.75
happy	0.86	0.86	0.93
sadness	0.50	0.17	0.33
surprise	0.88	0.75	0.81

By this table, we can find that the emotions of anger and sadness usually have higher probability to be misclassified. Neutral and Happy are more likely to be classified correctly.
Then we plotted the confusion matrix to evaluate the quality of the output of the classifier, obtained from the pooling model results.



From this plot, we can find anger emotions are easily classified as neutral emotions while sadness emotions can be classified as anger with relatively high probability. It might because of the face changing between anger and neutral is somewhat not so obvious. Basically, the difference is usually on the mouth movement. Happy with teeth showing and surprise with an O shape mouth will make these two emotions easily classified.
In a word, emotion recognition is very complex and hard, even for human eyes. In different context, the similar emotions can usually be interpreted as different meaning depending on the environment. Besides, the images we used were all taken at the front and very clear. The real world data won't be as good as this data. Further study needed to be conducted on the images which were not at the front nor clear, by using, such as, some feature descriptors procedure.

References

[1] Cohn-Kanade dataset
http://www.consortium.ri.cmu.edu/ckagree/

[2] Silhouette Value

https://en.wikipedia.org/wiki/Silhouette_(clustering)

[3] Kim, K. (1996). Face recognition using principle component analysis. In International conference on computer vision and Pattern recognition (pp. 586-591). [4] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classication. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1275{1284. ACM, 2009}.
[5]Sk-learn packages

http://scikit-learn.org/